

# Don’t Forget Your ABC’s: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems

Sarah E. Finch\* and James D. Finch\* and Jinho D. Choi

Department of Computer Science

Emory University

Atlanta, GA, USA

{sfillwo, jdfinch, jinho.choi}@emory.edu

## Abstract

Despite tremendous advancements in dialogue systems, stable evaluation still requires human judgments producing notoriously high-variance metrics due to their inherent subjectivity. Moreover, methods and labels in dialogue evaluation are not fully standardized, especially for open-domain chats, with a lack of work to compare and assess the validity of those approaches. The use of inconsistent evaluation can misinform the performance of a dialogue system, which becomes a major hurdle to enhance it. Thus, a dimensional evaluation of chat-oriented open-domain dialogue systems that reliably measures several aspects of dialogue capabilities is desired. This paper presents a novel human evaluation method to estimate the rates of many dialogue system behaviors. Our method is used to evaluate four state-of-the-art open-domain dialogue systems and compared with existing approaches. The analysis demonstrates that our behavior method is more suitable than alternative Likert-style or comparative approaches for dimensional evaluation of these systems.

## 1 Introduction

Recent work in human-computer chat has made remarkable progress. Multi-turn open-domain (MTOD) models are capable of holding engaging conversations with humans (Roller et al., 2021; Adiwardana et al., 2020). However, there remain a number of challenges facing MTOD chatbots such as hallucinations (Shuster et al., 2021), common-sense violations (Zhou et al., 2021), and consistency issues (Nie et al., 2021). A significant obstacle for research that addresses these challenges is the difficulty in formulating an appropriate evaluation methodology due to the inherent subjectivity in determining chat quality (van Miltenburg et al., 2021). Since existing automatic evaluation metrics have been shown to be biased measures of chat quality (Liu et al., 2016; Sai et al., 2019; Deriu

et al., 2022), evaluation using human judgments is standard, although the type of human judgments varies widely across works (Finch and Choi, 2020).

Overall, there are few works comparing and assessing the validity of various human evaluation methods. The result of this gap in the literature is that the relative sensitivity, interpretability, and importance of the metrics used to evaluate chat models are not well understood. A dimensional approach for evaluating chat models that measures different aspects of chat quality would surely aid progress (van Miltenburg et al., 2021). However, to our knowledge, no work has investigated the coverage of a comprehensive set of evaluation metrics. Consequently, existing chat model evaluation results provide an incomplete picture of the strengths and weaknesses of MTOD chatbots. This paper addresses these limitations of previous work through the following three contributions:

1. A novel, dimensional human evaluation method that measures the rate of chatbot behaviors impacting chat quality (Section 5).
2. A detailed validation of human evaluation methods, including likert scales and pairwise comparisons (Section 7).
3. A comprehensive evaluation of four MTOD chatbots using validated metrics (Section 8).

By presenting a detailed picture of MTOD chatbot performance and standard methods to evaluate them, we aid future work’s efforts to further understand and improve human-computer interaction. Our evaluation platform, analyses, and data are available at <https://github.com/emorynlp/ChatEvaluationPlatform>.

## 2 Chatbots

To evaluate the strengths and weaknesses of MTOD models, we select the chatbots for our study using a

\*Contributed equally to this work as first authors.

two-stage process: (1) a literature review to identify chatbot candidates, and (2) a pilot evaluation to select the final set of bots for our full study.

**Literature Review** To promote diversity among the selected chatbots, we focus our review on four popular themes of the human-computer chat: (1) Knowledge-grounded chat, (2) Empathetic chat, (3) Self-consistent chat, and (4) General open-domain chat with large pre-training resources like Reddit. Candidate chatbots are selected from each theme using the following criteria:

1. The bot must demonstrate state-of-the-art performance in a task related to the theme.<sup>1</sup>
2. The implementation must be provided.<sup>2</sup>
3. The response latency of the bot must be <10 seconds using modern GPU hardware.

This review yields the 6 chatbot candidates in Table 1: Blender-Decode (Nie et al., 2021), Blender2 (Weston and Shuster, 2021), BART-FiD-RAG (Shuster et al., 2021), Emora (Finch et al., 2020), DukeNet (Meng et al., 2020), and CEM (Sabour et al., 2022). Appendix A presents details of our literature review and selection process.

Model	Theme	N	Q	Pass
Blender-Decode	Consistency	10	4.1	✓
Blender2	General	10	3.8	✓
BART-FiD-RAG	Knowledge	10	3.5	✓
Emora	General	10	3.3	✓
DukeNet	Knowledge	9	1.9	✗
CEM	Empathy	12	1.1	✗

Table 1: The pilot results for 6 bots, showing the theme of the approach (**Theme**), the number of collected conversations (**N**), and the avg. dialogue-level Likert quality rating (**Q**). **Pass** denotes which models passed the verification criteria and were included in the full study.

**Chatbot Selection** A pilot evaluation using the 6 chatbot candidates is conducted in order to verify the multi-turn dialogue capability of the chatbot candidates. Appendix B provides details on the implementations of each chatbot candidate. 10 students majoring in Computer Science or Linguistics are invited to interact with randomly assigned chatbots in 3-5 text-based conversations,<sup>3</sup> each of

<sup>1</sup>Note that selection occurred in October 2021.

<sup>2</sup>We accepted either a trained English model or codebase with a fully-specified procedure to replicate the model.

<sup>3</sup>We use the web interface provided by ParlAI (Miller et al., 2017) hosted on our local webserver.

which consisted of 30 turns.<sup>4</sup> At the end of each conversation, students are asked to rate the quality from 1 (least) to 5 (most). Based on the pilot results (Table 1), DukeNet and CEM are excluded from our full study because they are unable to hold satisfying multi-turn conversations, despite their reasonable single-turn response generation capabilities. Appendix C shows example dialogues from these systems.

### 3 Conversation Collection

The conversation dataset used for the full study is collected using human interactors in a text-based conversation setting. 46 undergraduates are recruited as interactors. Each interactor is compensated with a \$5 gift card for every 6 conversations, and allowed to complete up to 18 conversations. Conversations are collected remotely using ParlAI’s interactive web interface, and links to the web interface are sent to each interactor with instructions to be completed within 2 weeks.

For each link, the interactor completes two conversations with a random pair of chatbots, for a minimum of 30 turns per conversation. We impose a similar open-ended, topic-free chatting environment to Adiwardana et al. (2020). Interactors are asked to rate 8 dimensions (Table 3) of each conversation after its completion on a 1-5 Likert scale, and to select the higher-quality conversation along the same 8 dimensions after each conversation-pair (ties allowed). Our final conversation dataset includes 400 human-bot dialogues (100 dialogues per chatbot), averaging 30.3 turns per dialogue (11.3 tokens per user turn).

### 4 Evaluation Methods

For a comprehensive evaluation of MOTD chatbots, a robust dimensional evaluation of their chat capabilities is crucial (van Miltenburg et al., 2021). To have confidence that any evaluation metric yields useful information, its interpretability and sensitivity require validation. In addition, it is important to verify that each evaluation metric provides distinct information relative to the others.

Several previous works propose sets of evaluation metrics that could be used for a dimensional evaluation but with insufficient analyses to validate them. Finch and Choi (2020) present an exhaustive set of metrics based on a literature survey of human

<sup>4</sup>A “turn” is defined as ONE message from a single interactor.

evaluation methods, but do not quantitatively validate its interpretability, sensitivity, or per-metric distinctness. Mehri and Eskenazi (2020a) present a set of Likert metrics and analyze their relationship to overall dialogue quality, but do not validate the sensitivity or distinctness of the individual metrics. Mehri and Eskenazi (2020b) present 5 Likert metrics and evaluate their coverage with respect to explaining single response quality, but do not validate their sensitivity or distinctness.

Similarly, some works look to identify common chatbot errors. Sanguinetti et al. (2020) and Higashinaka et al. (2021) present error taxonomies empirically grounded by error analyses, but do not present distinctness or sensitivity results for their error categories. See and Manning (2021) identify errors for one dialogue model and analyze the impact of each error on overall quality but do not attempt to verify the generalizability of their results.

Furthermore, various works propose novel evaluation methods with varying degrees of validation of the reliability and effectiveness of such methods. Deriu et al. (2020) present Spot the Bot, a pairwise evaluation approach that uses survival analysis to rank bots based on self-chats, but do not directly compare to alternative methodologies other than for cost. Sedoc and Ungar (2020) apply Item-Response Theory (IRT) (Lord and Novick, 2008) to pairwise comparison dialogue evaluation, by using a latent variable Bayesian model to estimate both the ability of the evaluated systems and the informativeness of inputs in the static evaluation set. Their analysis of the utility of IRT for dialogue evaluation does not include comparisons to existing approaches or a dimensional focus since they exclusively consider overall response quality. Ji et al. (2022) propose a continuous-scale method for evaluating multi-turn dialogue systems with quality control measures for mitigating artifacts from human annotators. They validate their proposed method on various dialogue dimensions using replication studies, a sensitivity analysis, and a correlation analysis between dimensions, although they explicitly acknowledge that their set of dimensions is not intended to be comprehensive. Phy et al. (2020) assert 3 dimensions (understandability, sensibleness, and likability) are sufficient for capturing the quality of a dialogue and validate their claims using agreement, correlation analysis, and distinctness analysis on human annotations of their dimensions, although they are not applied to multi-turn dialogues.

Two studies, Li et al. (2019a) and Smith et al. (2022), compare pairwise comparison and Likert evaluation methods via a sensitivity analysis. However, neither of them target a high-coverage set of dimensional metrics, as their studies were limited to 4 and 3 metrics respectively. Lee et al. (2020) also investigates pairwise evaluation using the Chat-Eval platform. However, this is not a multi-turn evaluation setup and it does not target a dimensional analysis since the comparisons are based exclusively on the overall quality of the responses.

	M	C	P	A	S	I	D
Finch and Choi (2020)	✓	✓					✓
Mehri and Eskenazi (2020a)	✓	✓	✓	✓			✓
Mehri and Eskenazi (2020b)	✓			✓			✓
Sanguinetti et al. (2020)	✓	✓		✓			
Higashinaka et al. (2021)	✓	✓	✓	✓			
See and Manning (2021)	✓				✓		✓
Deriu et al. (2020)	✓			✓	✓	✓	
Sedoc and Ungar (2020)				✓	✓		
Ji et al. (2022)	✓			✓	✓	✓	
Phy et al. (2020)					✓		✓
Li et al. (2019a)	✓		✓		✓		
Smith et al. (2022)	✓	✓		✓			
Lee et al. (2020)					✓	✓	
<b>This Work</b>	✓	✓	✓	✓	✓	✓	✓

Table 2: Recent studies of human evaluation metrics, in order of mention in Section 4. Multi-turn: investigates multi-turn response generation, Comprehensive: a set of metrics intended to explain dialogue quality, ComPared: compares alternative evaluation methods, Agreement: inter-annotator agreement, Sensitivity: validates metric sensitivity via statistical testing, Importance: relates evaluation metrics to overall dialogue quality, Distinctness: analyzes whether metrics provide distinct information about quality.

Overall, the relative validity of human evaluation metrics requires further investigation before a comprehensive and reliable dimensional evaluation of human-computer chat is achieved. Table 2 summarizes the goals and contributions of the previous evaluation works. Our study addresses all existing gaps by conducting a detailed validation study of 4 different human evaluation methods and a wide range of fine-grained metrics.

#### 4.1 Selected Methods

Four human evaluation methods are chosen for our study. Since MTOD chat model evaluation is our goal, any domain- or approach-specific methods or single-response evaluation methods providing

chatbots with a specific context are excluded.<sup>5</sup> We also focus on external human evaluation methods, where human evaluators judge conversations they do not participate in. Three of the selected methods represent popular approaches: Dialogue Likert, Turn Likert, and Comparative. The fourth method, ABC-Eval, is our novel evaluation approach.

Label	Dialogue Likert	Turn Likert	Comparative
Consistency	Con <sub>d</sub>	Con <sub>t</sub>	Con <sub>c</sub>
Emotion	Emo <sub>d</sub>	Emo <sub>t</sub>	Emo <sub>c</sub>
Understanding			
Engagingness	Eng <sub>d</sub>	Eng <sub>t</sub>	Eng <sub>c</sub>
Grammaticality	Gra <sub>d</sub>	Gra <sub>t</sub>	Gra <sub>c</sub>
Informativeness	Inf <sub>d</sub>	Inf <sub>t</sub>	Inf <sub>c</sub>
Quality	Qua <sub>d</sub>	Qua <sub>t</sub>	Qua <sub>c</sub>
Proactivity	Pro <sub>d</sub>	Pro <sub>t</sub>	Pro <sub>c</sub>
Relevance	Rel <sub>d</sub>	Rel <sub>t</sub>	Rel <sub>c</sub>

Table 3: The 8 labels for Likert and Comparative evaluations (taken from Finch and Choi (2020)), henceforth referred to using their abbreviations and colors.

**Dialogue Likert** Annotators provide dialogue-level ratings from 1 (least) to 5 (most) for the 8 labels shown in Table 3. We use the dimension set proposed in Finch and Choi (2020) which results from a detailed survey of characteristics used in chat evaluation and has better coverage than alternatives like the set used in ACUTE-Eval (Li et al., 2019a). Bot-level metrics are calculated as the mean rating across all bot dialogues.

**Turn Likert** Annotators provide turn-level ratings on the same scale and labels as those used for Dialogue Likert. The dialogue-level metric is measured as the mean rating of a single dialogue’s turns. The bot-level metric is calculated as the mean rating of all turns in all bot dialogues.

**Comparative** Annotators select the dialogue in which chatbot responses better fit a label definition from a side-by-side pair of dialogues, also using the labels in Table 3. A “neither” option is allowed, only for cases where the evaluator cannot distinguish which dialogue was a better fit. Bot-level metrics are calculated as bot pair win/tie/loss proportions between pairing of their dialogues.

**Behavior Classification: ABC-Eval** Annotators provide binary labels on the turn-level indicating

<sup>5</sup>We do not include a turn-level comparative evaluation because controlled comparisons require comparing turns with identical historical contexts which is not viable for real human-bot dialogues like those used in this work.

the presence or absence of a particular chat characteristic. The included chat characteristics are defined in Table 4. Dialogue-level metrics are calculated as the proportion of turns that display the characteristic of the dialogue. Bot-level metrics are calculated as the proportions of turns that display the characteristic over all bot dialogues. ABC-Eval is described in detail next in Section 5.

## 5 ABC-Eval Design

We hypothesize that binary turn-level behavior labels provide more reliable and informative metrics for quantifying fine-grained aspects of chat quality than alternative approaches such as Likert or Comparative scoring. Our novel method, the Annotation of Behaviors in Chat Evaluation (ABC-Eval), is developed in three stages: (1) collecting a set of behavior label candidates, (2) developing and piloting our annotation instructions and procedure, and (3) selecting a subset of behavior labels based on the validation study results in Section 7.

**Collecting Behavior Label Candidates** Based on a review of recent work in chat-oriented dialogue modeling and evaluation, we identify characteristics of chatbot responses relevant to conversation quality. These characteristics include those presented as error cases, evaluation metrics, or desirable response features. We then curate binarized definitions of these characteristics to create an initial set of behavior label candidates, which are revised through an iterative piloting and development process. Due to its high coverage of error categories, Higashinaka et al. (2021) is the primary source of inspiration for many of our behavior labels. However, we improve upon their presented taxonomies by considering additional labels based on characteristics of chat presented by other work, and by further refining their error categories to improve average Inter-Annotator Agreement (Section 7.1). Table 4 presents the final set and definitions of the 16 candidate behavior labels used in our full study, along with selected works from our review that inspired their inclusion. Appendix D details in full our development process.

**Annotation Procedure** The ABC-Eval procedure includes 16 binary behavior labels divided between 8 independent annotation tasks (Table 4). In each task, human evaluators are provided with definitions and examples of the behavior labels associated with that task and asked to annotate every

Label	Abbr.	Description	Inspired by
Uninterpretable	<b>!Int<sub>b</sub></b>	It is difficult to understand the intended meaning of part or all of the response.	1, 2, 3, 4, 5, 6
Antisocial	<b>!Soc<sub>b</sub></b>	The response is insulting, hateful, or excessively vulgar.	2, 7, 8, 9
Preference Info Life Info	<b>Pre<sub>b</sub></b> <b>Lif<sub>b</sub></b>	The response expresses the bot’s preferences, wishes, or values. The response shares information about the bot’s life or experiences.	10, 11
Empathetic Lack of Empathy	<b>Emp<sub>b</sub></b> <b>!Emp<sub>b</sub></b>	The response shows an understanding and reacts appropriately to someone’s emotions. The bot misunderstands or reacts inappropriately to someone’s emotions.	11, 12, 13
Commonsense Contradiction	<b>!Com<sub>b</sub></b>	The response misunderstands or contradicts common knowledge.	2, 14, 15, 16
Fact Usage <b>Fact Contradiction</b>	<b>Fac<sub>b</sub></b> <b>!Fac<sub>b</sub></b>	The response accurately incorporates encyclopedic or expert knowledge. The response hallucinates or inaccurately presents encyclopedic or expert knowledge.	1, 2, 11, 17, 18, 19, 20
Self Contradiction Partner Contradiction Redundant	<b>!Sel<sub>b</sub></b> <b>!Par<sub>b</sub></b> <b>Red<sub>b</sub></b>	The bot contradicts something it said earlier in the dialogue. The bot contradicts or misremembers something the user said earlier in the dialogue. The response inappropriately repeats information presented earlier in the dialogue.	2, 3, 6, 20, 21, 22, 23
Ignore Irrelevant Follow-up Topic Switch	<b>Ign<sub>b</sub></b> <b>!Rel<sub>b</sub></b> <b>Fol<sub>b</sub></b> <b>Top<sub>b</sub></b>	The response ignores what the user just said. The response interrupts the current topic of discussion by presenting unrelated information. The response explores, elaborates on, or asks about the ideas shared in the previous turn. The response introduces a new topic of conversation.	1, 2, 3, 6, 24

Table 4: The 16 behavior labels within ABC-Eval. Row separators denote evaluation task groupings. **Bold** indicates behavior labels kept in final set. [1] Gopalakrishnan et al. (2019), [2] Higashinaka et al. (2021), [3] Mehri and Eskenazi (2020a), [4] Mehri and Eskenazi (2020b), [5] Phy et al. (2020), [6] Sanguinetti et al. (2020), [7] Beattie et al. (2022), [8] Sun et al. (2022), [9] Xu et al. (2021), [10] Rashkin et al. (2021), [11] Smith et al. (2020), [12] Majumder et al. (2020), [13] Rashkin et al. (2019), [14] Zhong et al. (2021), [15] Zhou et al. (2021), [16] Zhou et al. (2022), [17] Gupta et al. (2022), [18] Honovich et al. (2021), [19] Santhanam et al. (2021), [20] Shuster et al. (2021), [21] Li et al. (2021), [22] Nie et al. (2021), [23] Welleck et al. (2019), [24] Xu et al. (2022) .

chatbot turn in a given human-chatbot conversation with each behavior label. Evaluators complete these tasks using a custom web application based on the ParlAI evaluation interface (Appendix G).

**Training and Screening** To improve annotation consistency and detect poorly performing evaluators, we develop automated training sessions each annotation task inspired by van Miltenburg et al. (2021). Each session consists of 3 conversations that evaluators annotate using an identical procedure and web interface to the corresponding task. The 3 conversations used for each session are hand-crafted by the authors to represent a variety of positive and negative examples of the behavior labels for the corresponding task (Appendix D). The gold annotations for each training conversation are hidden from evaluators during the annotation; however, after completing each training conversation, any disagreements between the evaluator’s annotations and gold labels are displayed along with an explanation to help the evaluator improve. We use the evaluator’s performance on the third conversation of each training session to screen evaluators, where performance is measured by the number of turns where their annotations disagree with gold labels. Evaluators are eligible to complete the work on a task if they make mistakes on fewer than 2 turns for the antisociality and uninterpretability tasks, or on fewer than 3 turns for the other 6 tasks.

## 6 Evaluation Study

Our full study consists of the collection of 40 labels per conversation. This collection was split into 18 independent evaluation tasks as follows:

- 8 ABC-Eval tasks, each composed of 1 to 4 labels as denoted by groupings in Table 4
- 1 Dialogue Likert task, composed of all 8 labels from Table 3 completed in random order
- 8 Turn Likert tasks, each composed of 1 label from Table 3
- 1 Comparative task, composed of all 8 labels from Table 3 completed in random order

The 18 evaluation tasks are posted on SurgeHQ’s annotation platform<sup>6</sup> to be completed by dedicated remote workers (Surgers) with experience in NLP annotation. Each time an evaluator connects to one of our tasks, they are assigned a randomly selected conversation to annotate. We are allocated a group of 125 Surgers, chosen by a SurgeHQ employee based on high annotation performance on past projects. Evaluators are compensated per annotated conversation per task, at an estimated rate of \$20/hr<sup>7</sup>. We allow evaluators to annotate up to 60 conversations per task.

<sup>6</sup><https://www.surgehq.ai>; Appx. E details annotator selection.

<sup>7</sup>Per-task payment rates provided in Appendix F.

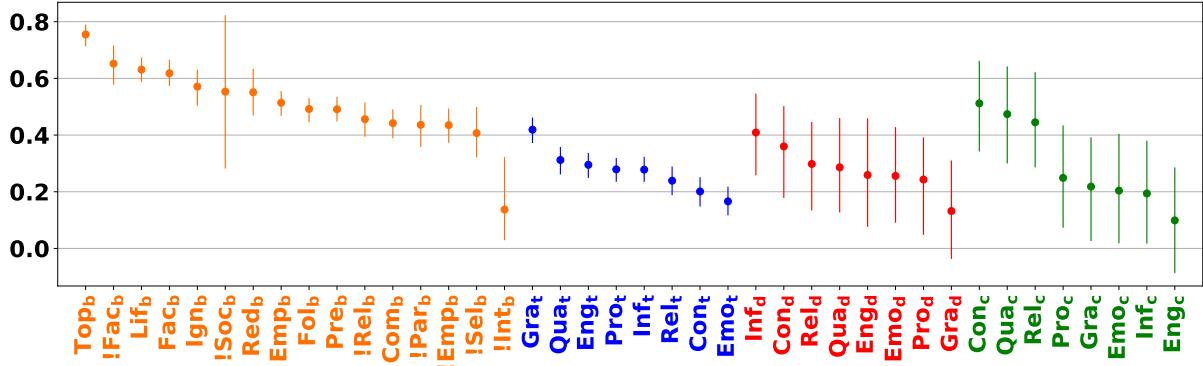


Figure 1: IAA (Krippendorff’s alpha) for all metrics. Error bars denote 95% bootstrap<sup>9</sup> confidence intervals.<sup>10</sup>

Our final evaluation dataset consists of 400 conversations, each with results for all 40 labels.<sup>8</sup> Additionally, a randomly-selected subset of 100 conversations (and 50 of the conversation pairs) is evaluated a second time by a different Surger in order to measure IAA.

## 7 Metric Analysis

### 7.1 Interpretability

We measure the reliability of interpreting each metric’s annotation instructions by calculating IAA using our set of 100 double-annotated conversations (Figure 1). High agreement between annotators demonstrates that different people can reliably come to the same conclusions about how a metric’s definition applies to each chatbot response.

Our results suggest that the definitions of most ABC-Eval metrics can be interpreted more reliably than the definitions of most Dialogue Likert, Turn Likert, and Dialogue Comparison metrics. Likert-style and comparison-style annotations appear to have similar interpretability, although *Qua<sub>c</sub>* was a notable exception that produced higher agreement than *Qua<sub>d</sub>*.

### 7.2 Importance

The importance of each metric is estimated by a predictive validity analysis that measures the extent, to which the metric can predict conversation quality (Figure 2). We use *Qua<sub>d</sub>* and *Qua<sub>c</sub>* from interactors that participated in the conversations (Section 3) to avoid cases where the same evaluator produced the quality label and explanatory metric.

<sup>8</sup>Only 192 of our 200 dialogue pairs were evaluated with Comparative labels due to a collection mistake

<sup>9</sup>Bias-corrected and accelerated confidence intervals with  $k = 10,000$  Monte Carlo case resamples.

<sup>10</sup>*!Soc<sub>b</sub>* and *!Int<sub>b</sub>*’s confidence intervals are largely due to a low rate of positive examples (see Figure 4).

The predictive validity of each metric was measured by fitting univariate linear or logistic regression models to predict *Qua<sub>d</sub>* or *Qua<sub>c</sub>*, respectively.

*Qua<sub>c</sub>* was represented as a binary encoding, where 0 and 1 represent choosing the first and second conversation, respectively. We excluded any conversation pairs in which the interactor could not distinguish a difference in quality between conversations, and fitted models on the remaining set of 184 conversations. To use non-comparative predictors for predicting *Qua<sub>c</sub>*, the difference in metric value between each pair of conversations was used.

Our results suggest that dialogue quality is substantially related to emotional understanding metrics (*Emo*, *Emp<sub>b</sub>*, *!Emp<sub>b</sub>*), relevance-related metrics (*Rel*, *!Rel<sub>b</sub>*, *Ign<sub>b</sub>*), and consistency metrics (*Con*, *!Sel<sub>b</sub>*, *Red<sub>b</sub>*, *!Par<sub>b</sub>*). Within these metric groupings, ABC-Eval metrics were overall more predictive of quality than their Likert or comparative analogs, while comparative metrics were least predictive of quality. Chatbots’ ability to express knowledge (*Inf*, *Fac<sub>b</sub>*, *!Fac<sub>b</sub>*, *Lif<sub>b</sub>*, *Pref<sub>b</sub>*) was an overall poor predictor of quality; however, commonsense knowledge errors (*!Com<sub>b</sub>*) was one of the strongest predictors.

### 7.3 Sensitivity

We investigate the sensitivity of each metric using two analyses. First, we use the fitness of the univariate regression models described in the previous section as one source of evidence for metric sensitivity, since a metric must be sufficiently sensitive in order to distinguish conversations of low and high quality. Second, we follow Li et al. (2019a) and run hypothesis tests to count the number of statistically significant differences each metric is able to detect between the 6 pairings of our 4 chatbots (Table 5). To make results comparable, we

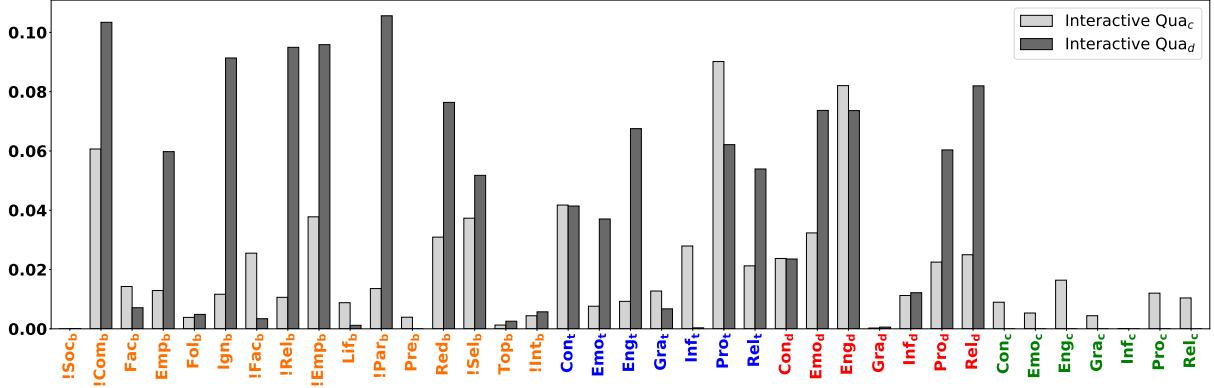


Figure 2: Extent to which each evaluation metric can explain variance in conversation quality by fitting a univariate regression model ( $R^2$  for predicting  $Qua_d$  using linear regression, McFadden’s Pseudo- $R^2$  for predicting  $Qua_c$  using logistic regression). Comparative metrics cannot predict  $Qua_d$  so only results for variance of  $Qua_c$  are shown.

$\alpha$	!Soc <sub>b</sub>	!Comb	Fac <sub>b</sub>	Emp <sub>b</sub>	Fol <sub>b</sub>	Ign <sub>b</sub>	!Fac <sub>b</sub>	!Rel <sub>b</sub>	!Emp <sub>b</sub>	Lif <sub>b</sub>	!Par <sub>b</sub>	Pre <sub>b</sub>	Red <sub>b</sub>	!Sel <sub>b</sub>	Top <sub>b</sub>	Int <sub>b</sub>	Con	Emot	Engt	Grat	Inf <sub>t</sub>	Prot	Rel <sub>t</sub>	Cond	Emo <sub>d</sub>	Eng <sub>d</sub>	Grad	Inf <sub>d</sub>	Prod	Rel <sub>d</sub>	Con <sub>c</sub>	Emo <sub>c</sub>	Eng <sub>c</sub>	Grac	Inf <sub>c</sub>	Proc	Rel <sub>c</sub>			
0.01	0	1	4	3	5	1	4	2	4	2	0	4	2	5	5	2	1	1	2	4	4	4	4	2	1	2	1	0	0	3	1	0	1	0	0	0				
0.05	0	2	5	3	5	2	6	2	5	2	1	4	3	5	5	3	2	1	3	5	4	4	4	3	2	3	2	3	2	3	1	2	1	0	0	0				
0.1	1	3	5	3	5	2	6	2	5	3	1	4	3	5	5	3	2	3	3	5	4	4	4	4	2	4	2	5	3	3	3	2	2	0	1	0	2	2	1	2

Table 5: The number of statistically significant differences detected by each metric when comparing bot-pairs using z-tests of proportions (ABC-Eval), t-tests (Turn Likert and Dialogue Likert), and sign tests (Comparative) at three significance thresholds.

downsample the conversations used for hypothesis testing to 32 conversations per bot for our Dialogue Likert, Turn Likert, and ABC-Eval metrics to match the 32 conversation-pairs per bot-pair produced by our Comparative evaluation.

Our results show that the Likert evaluations were more sensitive than the Comparative evaluation for most labels. ABC-Eval metrics have a wide range of sensitivity, with some ABC-Eval metrics appearing to be more sensitive analogs of similar likert metrics. For example, the results suggest that *!Sel<sub>b</sub>* and *Red<sub>b</sub>* are more sensitive than *Con*, that *Fac<sub>b</sub>* and *!Fac<sub>b</sub>* are more sensitive than *Inf*, and that *Emp<sub>b</sub>* and *!Emp<sub>b</sub>* are more sensitive than *Emo*. On the other hand, the likert-style *Rel* metric shows similar or slightly superior sensitivity compared to the analogous *Ign* and *!Rel* behavior metrics.

#### 7.4 Coverage & Distinctness

We investigate the coverage and distinctness of our metrics via incremental validity analysis. For this analysis, we perform backwards stepwise regression that determines (1) the ability of an evaluation method as a whole to explain conversation quality, and (2) whether each metric contributes distinct information about quality above and beyond other metrics (Figure 3). Specifically, we fit a multivariate regression model for each of our 4 evaluation methods. These models are fit similarly to those

presented in Section 7.2, but include all non-quality metrics within an evaluation method as predictors. Then, we remove predictors from each model one at a time based on a beam search ( $k=100$ ) of which removed predictor results in the smallest decrease in model fitness (adjusted  $R^2$  or adjusted pseudo- $R^2$ ). We perform this stepwise regression analysis twice to predict both  $Qua_d$  and  $Qua_c$  given by interactors, similar to our analysis in Section 7.2. Our results suggest that ABC-Eval has overall better coverage than other evaluation methods for explaining conversation quality. Furthermore, most ABC-Eval metrics that have a strong relationship with conversation quality appear to be appropriately distinct in the information they provide, especially *!Emp<sub>b</sub>*, *!Sel<sub>b</sub>*, *Red<sub>b</sub>*, *!Rel<sub>b</sub>*, *Emp<sub>b</sub>*, *!Com<sub>b</sub>*, and *Ign<sub>b</sub>*. Similar distinctness can also be seen in Turn Likert metrics, whereas dialogue-level metrics show relatively low distinctness.

#### 7.5 Final ABC-Eval Metrics

Given the results of our metric analysis, we select the final set of ABC-Eval metrics bolded in Table 4. In our analyses, this final set had better interpretability (Section 7.1), a wider coverage of distinct characteristics of chat that impact quality (Section 7.2 and Section 7.4), and overall higher measurement sensitivity (Section 7.3) than alternative evaluation methods. Furthermore, the fi-

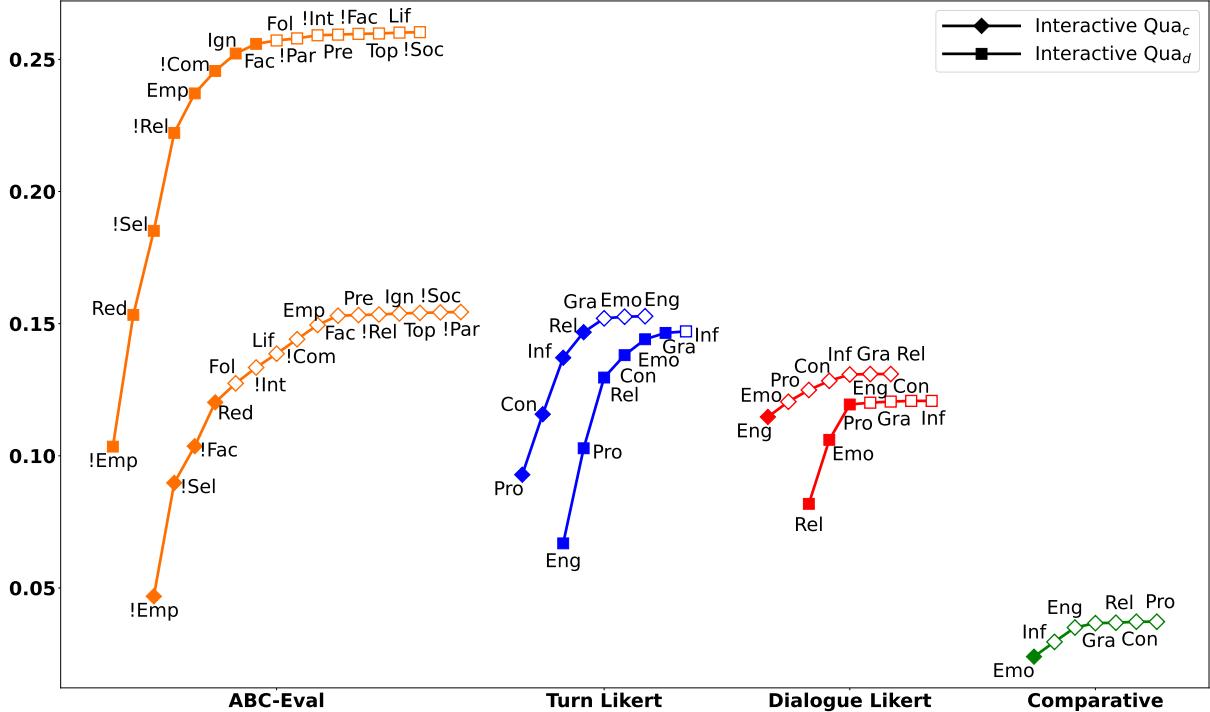


Figure 3: Incremental validity of metrics within 4 evaluation methods, obtained using backwards stepwise regression. Points represent the extent to which a model can explain variance in quality ( $R^2$  for predicting  $Qua_d$  with a linear model, McFadden’s pseudo- $R^2$  for predicting  $Qua_c$  with a logistic model) using all metrics on the same line and to the left as predictors. Filled marker symbols denote steps where the model’s predictors all contributed positively to adjusted  $R^2$  or adjusted pseudo- $R^2$  values; otherwise, marker symbol is unfilled. Comparative metrics cannot be used to predict  $Qua_d$  so only results for explaining variance of  $Qua_c$  are shown.

nal ABC-Eval metrics are less costly<sup>11</sup> (a median of 15.2 min/dialogue) to collect than Turn Likert metrics (19.9 min/dialogue). Although dialogue-level evaluations are least costly (2.8 min/dialogue for Dialogue Likert, 4.4 min/dialogue for Comparative), our results suggest that dialogue-level judgements may be ill-suited for dimensional evaluation, since the dialogue-level metrics we tested had worse coverage and distinctness (Section 7.4).

## 8 Chatbot Evaluation

To evaluate the strengths and weaknesses of our 4 selected chatbots, we present results for the 400 collected conversations across all ABC-Eval metrics (Figure 4 and Figure 5), Likert Dialogue metrics (Figure 6), Likert Turn metrics (Figure 8), and Comparative metrics (Figure 7). We focus our discussion on the final set of ABC-Eval metrics since they performed best in our metric analysis.

The results highlight the notable recent progress in human-computer chat, as the vast majority of chatbot turns are interpretable, relevant responses to the dialogue context. Less than 1% of re-

sponses have interpretability issues, and Blender2 and BART-FiD-RAG each achieve a relevant response rate of nearly 90%. Blender2 specifically is also able to incorporate factual knowledge into about 20% of its responses while hallucinating factual information at a remarkably low rate, less than 1%. Furthermore, the chatbots almost never produce responses with offensive language.<sup>12</sup> The chatbots also show a high rate of emotional understanding, with 40% of their responses containing emotionally-appropriate reactions to the user.

Despite these strengths, our results also show several clear directions for improvement. Common-sense violations are present in about 15-20% of the bots’ responses. Consistency issues are prevalent across all bots: self-contradictions, partner contradictions, and redundancies appear in about 5% of the bots’ responses overall. Also, all chatbots have a substantial rate of violating natural dialogue structure: about 10% of responses are judged as ignoring the user, and depending on the chatbot,

<sup>11</sup>See Appendix F for detailed cost results.

<sup>12</sup>Note that our experiments are conducted with cooperative human interactors. Chatbots similar to those we test have been shown to reliably produce offensive language when responding to provocative inputs (Dinan et al., 2022).

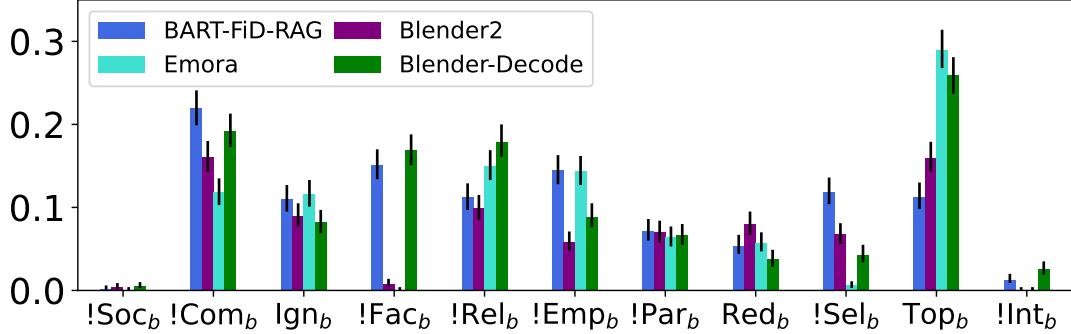


Figure 4: Proportions of turns expressing undesirable behaviors, with 95% Wilson score confidence intervals.

around 10-20% of responses include irrelevant contributions to the dialogue. Additionally, 5-15% of the chatbots’ responses show a lack of empathy or other emotional misunderstandings. The reality of these observed rates of problematic behaviors is that, in most 30-turn conversations with these chatbots, a human interactor is likely to experience several issues that impact conversation quality.

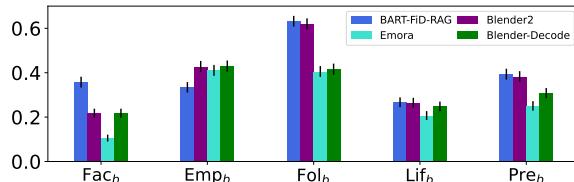


Figure 5: Proportions of turns expressing desirable behaviors, with 95% Wilson score confidence intervals.

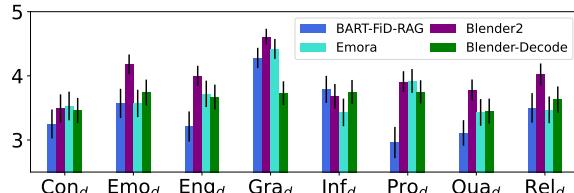


Figure 6: Average Dialogue Likert ratings of the conversations, with 95% Student’s t confidence intervals.

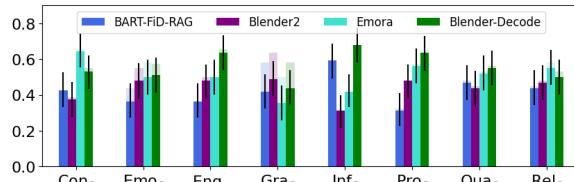


Figure 7: Win/tie rates of each bot vs all other bots, with 95% Wilson score confidence intervals for win proportion. Transparent segments denote tie rates.

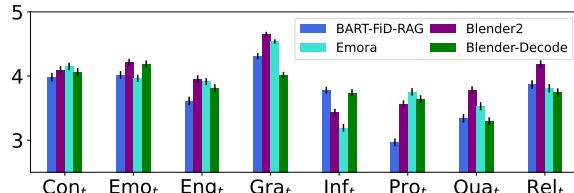


Figure 8: Average Turn Likert ratings of the conversations, with 95% Student’s t confidence intervals.

## 9 Conclusion

As illustrated here, dialogue quality is a complex construct with many dimensions. Depending on the approach, dialogue systems can have markedly different weaknesses among these quality dimensions. Our research highlights several outstanding challenges, especially regarding the relevance, consistency, common-sensibility, and emotional understanding of chat model responses. Our analyses not only demonstrate that these four dimensions have a high impact on conversation quality, but also that current chatbots have substantial response error rates in these areas.

To efficiently address the challenges facing open-domain dialogue models, we need a reliable, dimensional evaluation method; however, our results show that popular evaluations such as dialogue-level Likert and comparative methods may not be suitable. The presented ABC-Eval serves as a promising alternative in this direction. Although the popular dialogue-level likert evaluation method may be the most cost-effective and robust method for measuring overall dialogue quality, we recommend that researchers additionally use the final set of ABC-Eval metrics, or a subset relevant to their scientific goals, to evaluate specific strengths and weaknesses of new chat models. Overall, we hope future work can use insights from our study to make better-informed decisions about which evaluation method to use, and to tackle the challenges facing current chatbots.

## 10 Limitations

There are several characteristics of the presented analyses that limit the scope of conclusions that can be drawn. We discuss how each of these limitations affect the takeaways of our results below.

**Number of Chatbots** The generalizability of our metric analysis results (Section 7) is constrained by

the fact that we were only able to include conversations from 4 chatbots in our analyses. We did our best to choose chatbots representative of the field and seem to have selected a fairly diverse group of models (Section 8). However, it is possible that not all results we found in our metric analyses will generalize when evaluating other chat models. One possible example is the number of partner contradictions we observed among our 4 chatbots (Figure 4), which may be similar by coincidence. If other chatbot models indeed differ more substantially in partner contradiction rates, our sensitivity metric analysis may have underestimated the sensitivity of our partner contradiction metric (Section 7.3). In general, including a larger number of chatbots in a metric analysis will improve the chance that its results will apply to new chatbot models. Future work that performs metric analyses like those we presented, but with different chatbots than the 4 selected in this work, would aid further analysis of our results’ generalizability.

**Use of Surgeons as Evaluators** We perform our analyses using only a single evaluator group (Surgeons). This choice of evaluator group does not harm the replicability of our methods, as other researchers have access to use of SurgeHQ or similar third-party annotation companies. However, several other evaluator groups are more popularly used for chat model evaluation, such as university students and Amazon Mechanical Turkers (MTurkers). We attempted to carry out our study with three evaluator groups (see Appendix E for details), but were unable to proceed with student and MTurker evaluator groups due to time constraints. Consequently, it is unclear to what extent our metric analysis results will generalize to other choices of evaluator.

**Number of Collected Conversations** As with any study involving a sampling procedure, resource constraints limit the number of collected samples, which in turn limits the statistical power of the study’s analyses. Our study included 400 conversations, which provided more than adequate statistical power for most of our analyses. For example, our investigation of each metric’s predictive validity (Section 7.2) relied on a simple linear regression analyses. At a significance level of  $\alpha=0.05$ , our 400 conversation samples would yield a statistical power of  $1-\beta=0.80$  to detect effect sizes of  $f^2=0.14^2$  by F-test for each metric’s regression. However, our analyses with the weakest statistical

power are our dialogue-level analyses that compare bots with only 100 samples per bot. At 100 samples per bot, and assuming a standard deviation of 1.0 Likert points,<sup>13</sup> a two-tailed t-test of mean Dialogue Likert rating would have a statistical power of  $1-\beta=0.80$  to detect differences of an effect size of Cohen’s  $d=0.40$ . This is still a reasonable amount of statistical power, but leaves room for our study to produce inconclusive results when the true differences between chatbots are small.

## 11 Ethics Statement

The presented work aims towards improving the scientific methodology of chat model evaluation. To this end, we present a battery of analyses comparing several aspects of metric validity for four different evaluation methods (Section 7). Our results allow other researchers in the field to make better-informed decisions regarding appropriate evaluation methodology in human-computer chat. To ensure replicability of our methods we publicly release the annotation software and chatbot implementations used to collect our conversation and evaluation data. Additionally, we provide full transparency in our analyses by releasing the code for all our presented analyses. Finally, to aid future research efforts in human-computer chat modelling and evaluation, we release an anonymized version of our conversation and evaluation data.

One ethical consideration involved in our work involved managing human workers in our data collection processes. All worker participation in our study was voluntary and involved zero subjective screening processes, with a complete description of worker tasks, workload, and timeframe provided before work was assigned. Workers could opt out of our study at any time for any reason. As compensation for work completed, we targeted a compensation rate of \$10/hour for student<sup>14</sup> and Amazon Mechanical Turk workers, and a rate of \$20/hour for Surgeons. We compensated on a per-task-completed basis to ensure timely completion of work, but verified that target hourly rates were reasonably approximated throughout the course of the study by measuring workers’ median task completion times (see Appendix F for details). These measures ensured that all human work in our study

---

<sup>13</sup>Smith et al. (2022) reports standard deviations of Likert metrics between 0.8 and 1.3

<sup>14</sup>Students’ compensation is given as an Amazon Gift Card for convenience; students are informed of this prior to any work being completed

was fair, transparent, and mutually-beneficial.

Other ethical considerations arise in our study’s conversation collection. Unlike the collection of evaluation or annotation data, collecting interactive conversation data from human-computer interaction poses a small but meaningful risk that sensitive, damaging, or personally identifying information could get collected. We mitigated this risk in three ways. First, students were notified in multiple email communications and before each conversation that their conversations with our chatbots would be publicly released. Included in these notices was the instruction to refrain from releasing any personally identifiable or damaging information. Our instructions suggest that students fabricate personal information at any time during the conversations if it would make them feel more comfortable. Second, we hand-checked all 400 conversations to ensure the non-presence of any sensitive information. Third, we anonymize all data before public release. Our study’s collection and analysis of conversation data did not investigate interactors as human subjects, and we did not seek institutional review board approval.

Finally, there is a concern in our study about the potential of the chatbots to respond to student interactors with toxic, insensitive, or vulgar language. The data-driven nature of some of our evaluated chat models means the chatbots are prone to reflecting any biases, toxicity, and vulgarity present in the training data (see [Dinan et al. \(2022\)](#) for a quantitative analysis). A high rate of antisocial behaviors among our evaluated models could potentially make human interactors’ experience talking with the bots quite uncomfortable, and would poorly reflect on the research field’s potential for social good. To mitigate this risk, the authors extensively hand-tested all evaluated chat models, as well as conducting a pilot evaluation among the authors’ lab group. As confirmed further in our results (Section 8), our chatbots exhibited negligible rates of antisocial behavior.

## 12 Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon. In addition, thank you to Bradley Webb, Scott Heiner, and the rest of the SurgeHQ team for their guidance in run-

ning our annotation projects on their platform. We are also grateful to our colleagues at Emory for their participation in piloting the bots and refining the annotation interfaces. Lastly, a thank you to our reviewers for their helpful feedback.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a Human-like Open-Domain Chatbot](#). ArXiv:2001.09977 [cs, stat].
- Hedin Beattie, Lanier Watkins, William H. Robinson, Aviel Rubin, and Shari Watkins. 2022. [Measuring and Mitigating Bias in AI-Chatbots](#). In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, pages 117–123.
- Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. [Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.
- Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. [Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.
- Jan Deriu, Don Tuggener, Pius Von Däniken, and Mark Cieliebak. 2022. [Probing the Robustness of Trained Metrics for Conversational Dialogue Systems](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 750–761, Dublin, Ireland. Association for Computational Linguistics.
- Emily Dinan, Gavin Abercrombie, A Bergman, Shannon L Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. Safetykit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of Wikipedia: Knowledge-Powered Conversational Agents](#). In *Proceedings of the International Conference on Learning Representations*.
- Sarah E. Finch and Jinho D. Choi. 2020. [Towards Unified Dialogue System Evaluation: A Comprehensive](#)

**Analysis of Current Evaluation Protocols.** In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. 2020. **Emora: An Inquisitive Social Chatbot Who Cares For You**.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. **Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations**. In *Interspeech 2019*, pages 1891–1895. ISCA.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. **DialFact: A Benchmark for Fact-Checking in Dialogue**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. **Integrated taxonomy of errors in chat-oriented dialogue systems**. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. **Q2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. **Achieving Reliable Human Assessment of Open-Domain Dialogue Systems**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. **Dense passage retrieval for open-domain question answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal,

Dilek Hakkani-Tur, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. **Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize**.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. **Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. **Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. **Internet-Augmented Dialogue Generation**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Seolhwa Lee, Heuiseok Lim, and João Sedoc. 2020. **An Evaluation Protocol for Generative Conversational Systems**. ArXiv:2010.12741 [cs].

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. **Don’t Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019a. **ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons**.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019b. **Incremental Transformer with Deliberation Decoder for Document Grounded Conversations**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. **Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. **How NOT To Evaluate Your Dialogue System: An**

- Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Frederic M Lord and Melvin R Novick. 2008. *Statistical theories of mental test scores*. IAP.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. **MIME: MIMicking Emotions for Empathetic Response Generation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.
- Shikib Mehri and Maxine Eskenazi. 2020a. **Unsupervised Evaluation of Interactive Dialog with Di-aloGPT**. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.
- Shikib Mehri and Maxine Eskenazi. 2020b. **USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.
- Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. **DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation**. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’20, pages 1151–1160, New York, NY, USA. Association for Computing Machinery.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. **ParlAI: A Dialog Research Software Platform**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. **I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.
- Timo Partala and Veikko Surakka. 2004. **The effects of affective interventions in human–computer interaction**. *Interacting with Computers*, 16(2):295–309.
- Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. **Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Helmut Prendinger and Mitsuru Ishizuka. 2005. **The Empathic Companion: A Character-Based Interface That Addresses Users’ Affective States**. *Applied Artificial Intelligence*, 19:267–285.
- Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. **Conversational AI: The Science Behind the Alexa Prize**.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. **Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. **Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. **Recipes for Building an Open-Domain Chatbot**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. **CEM: Commonsense-Aware Empathetic Response Generation**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237. Number: 10.
- Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. **Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6220–6227. Number: 01.
- Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalariandi, Dario Mana, and Rossana Simeoni. 2020. **Annotating Errors and Emotions in Human-Chatbot Interactions in Italian**. In *Proceedings of the 14th Linguistic Annotation Workshop*,

- pages 148–159, Barcelona, Spain. Association for Computational Linguistics.
- Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. **Rome was built in 1776: A Case Study on Factual Correctness in Knowledge-Grounded Response Generation**. ArXiv:2110.05456 [cs].
- João Sedoc and Lyle Ungar. 2020. **Item Response Theory for Efficient Human Evaluation of Chatbots**. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.
- Abigail See and Christopher Manning. 2021. **Understanding and predicting user dissatisfaction in a neural generative chatbot**. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. **Retrieval Augmentation Reduces Hallucination in Conversation**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. **Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents**. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. **Can You Put it All Together: Evaluating Conversational Agents’ Ability to Blend Skills**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.
- Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. **BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.
- Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. **Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online. Association for Computational Linguistics.
- Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. **Generating Responses with a Specific Emotion in Dialog**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.
- Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. **On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark**. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. **Underreporting of errors in NLG output, and what to do about it**. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. **Dialogue Natural Language Inference**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.
- Jason Weston and Kurt Shuster. 2021. **Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet**.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. **Bot-Adversarial Dialogue for Safe Conversational Agents**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. **Beyond Goldfish Memory: Long-Term Open-Domain Conversation**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.
- Hao-Tong Ye, Kai-Lin Lo, Shang-Yu Su, and Yun-Nung Chen. 2020. **Knowledge-Grounded Response Generation with Deep Attentional Latent-Variable Model**. *Computer Speech & Language*, 63:101069.
- Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. **CoLV: A Collaborative Latent Variable Model for Knowledge-Grounded Dialogue Generation**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. **DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. **CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14577–14585. Number: 16.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. **Commonsense-Focused Dialogues for Response Generation: An Empirical Study**. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. **Think Before You Speak: Explicitly Generating Implicit Commonsense Knowledge for Response Generation**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

## A Chatbot Selection Details

This appendix discusses the details of our literature review for each of the four chosen research themes.

**General** Our work focuses primarily on open-domain chat. Large-scale language modeling using dialogue-like pretraining data can produce surprisingly human-like conversation on virtually any popular conversation topic (Roller et al., 2021). Of these approaches, we chose Blender2 (Weston and Shuster, 2021), which reportedly outperformed its previous iteration Blender (Roller et al., 2021) who had surpassed DialoGPT (Zhang et al., 2020) and Meena (Adiwardana et al., 2020).

There had been also several chatbots produced by the Amazon Alexa Prize Socialbot Grand Challenge (Ram et al., 2018) focusing on general, open-domain chat, most of which incorporate rule-based methods to ensure interesting and consistent responses (Khatri et al., 2018). Since these chatbots performed well in practice but lack comparison to SOTA data-driven models, we selected the bot with the all-time highest final score, Emora<sup>15</sup> (Finch et al., 2020), as one of our candidates.

**Knowledge** Grounding chat with supplementary knowledge resources is a common way to improve engagingness and control the topic of conversation (Li et al., 2019b; Ye et al., 2020). ColV (Zhan et al., 2021) achieved SOTA performance in knowledge-grounded dialogue response generation on the popular WoW dataset (Dinan et al., 2019); however, no implementation was publicly available. DukeNet (Meng et al., 2020) and PIPM (Chen et al., 2020) report next-best performance in this task. DukeNet’s implementation was available while PIPM’s was not, therefore we selected DukeNet as a candidate.

BART-FiD-RAG also reported compelling performance for knowledge-grounded chat (Shuster et al., 2021), but did not compare to other SOTA models we identified. Since BART-FiD-RAG’s inclusion in ParlAI provided easy replication, we included it in our bot pilot.

**Consistency** Improving consistency of chatbot responses is noted as a challenge and addressed in several works (Welleck et al., 2019; Nie et al., 2021; Li et al., 2021). DECODE (Nie et al., 2021) reported SOTA performance for general inconsistency avoidance, improving upon an approach that used unlikelihood training with dialogue natural

language inference data (Li et al., 2020). Note that there were several works focusing specifically on persona consistency (Song et al., 2020; Kim et al., 2020; Song et al., 2021), which we did not consider due to their narrower contradiction scope.

**Empathy** Several works demonstrated the importance of emotional understanding in chat (Partala and Surakka, 2004; Prendinger and Ishizuka, 2005; Kim et al., 2021; Sabour et al., 2022). To provide contrast with our knowledge-grounded candidates, we selected CEM (Sabour et al., 2022), which reported SOTA results in empathetic response generation. Many related works investigated controllable emotional response generation (Song et al., 2019; Zhong et al., 2021), but we did not consider models requiring an emotion label as input.

## B Chatbot Implementation Details

For each selected candidate model, a brief overview of the implementation details required to use them as interactive models in this work is below:

**Emora** We implement a ParlAI agent using the interactive chatting mode provided for the Emora system (Finch et al., 2020).

**BART-FiD-RAG** An interactive chatting mode for BART-FiDRAG is provided through ParlAI. We use the default arguments, except for compressed-indexer-nprobe and beam-context-block-ngram which we set to 128 and 3, respectively, following internal testing by the authors before the pilot.

**Blender2** An interactive chatting mode for Blender2<sup>16</sup> is provided through ParlAI; however, the internet-search augmented functionality as described in Komeili et al. (2022) is not provided. Although there is an available alternative internet search server<sup>17</sup> that is compatible with Blender2, it increased the response latency of Blender2 such that it no longer satisfied our response latency requirement. As a result, we use the 3B parameter version of Blender2 that does not include the internet search functionality.

**Blender-Decode** We implement a custom agent in the Parlai framework following the best-performing setup of Nie et al. (2021) of finetuning a consistency classification model and using it to

<sup>16</sup><https://parl.ai/projects/blenderbot2/>

<sup>17</sup>[https://github.com/JulesGM/ParlAI\\_SearchEngine](https://github.com/JulesGM/ParlAI_SearchEngine)

<sup>15</sup>[https://github.com/emora-chat/emora\\_ap3\\_parlai](https://github.com/emora-chat/emora_ap3_parlai)

rerank generated responses from Blender (Roller et al., 2021) such that the least contradictory response is selected.

**DukeNet** The published code for DukeNet is missing an interactive chatting mode, although a trained model on the multi-topic Wizard of Wikipedia (WoW) (Dinan et al., 2019) is provided. We modified the published code to support an interactive chatting mode. However, one major difference is that the relevant knowledge documents need to be retrieved in interactive chatting mode, whereas they are predetermined in the training and evaluation environments. To remedy this, we use a trained Dense Passage Retrieval model (Karpukhin et al., 2020) provided through Sentence-Transformers<sup>18</sup> to select knowledge at each dialogue turn. We implement a ParlAI agent using the available DukeNet model and our custom interactive chatting mode.

**CEM** The published code for CEM is missing both an interactive chatting mode and a trained model. We trained a CEM model using the default arguments from the repository and modified the published code to support an interactive chatting mode. We implement a ParlAI agent using this trained model and our custom interactive chatting mode.

<b>User</b>	Hi!
<b>Bot</b>	What is happening?
<b>User</b>	I am just getting my day started. Finished breakfast and booted up my computer to begin working. What about you?
<b>Bot</b>	That is great! I wish you the best of luck!
<b>User</b>	Yeah, I think it will be a good day and I am excited for it. What are you doing today?
<b>Bot</b>	I am sure you will be fine!
<b>User</b>	I'm not too worried about it.
<b>Bot</b>	I am so happy for you!
<b>User</b>	So, what do you do for work?
<b>Bot</b>	I am sure you will do great!

Table 6: Chatbot pilot conversation: CEM

<b>User</b>	Hi!
<b>Bot</b>	Hello, how are you?
<b>User</b>	I am pretty good what about you
<b>Bot</b>	I'm doing well, thanks for asking. How's the weather in your neck of the woods?
<b>User</b>	pretty warm
<b>Bot</b>	That's good. Do you have any plans for the weekend? I'm thinking of going to the beach.
<b>User</b>	cool I have a gathering with my church friends
<b>Bot</b>	That sounds like a lot of fun. I haven't been to church in a long time. How long have you been going to church?
<b>User</b>	many years since I was very young
<b>Bot</b>	That's cool. I don't think I've ever been to a church service in my life.
<b>User</b>	what do you do for fun
<b>Bot</b>	I like to watch movies and play video games. What do you like to do in your free time?

Table 7: Chatbot pilot conversation: Blender2

## C Bot Pilot Examples

Tables 6 and 7 show two of our chatbot pilot (Section A) conversations, one from CEM (Sabour et al., 2022) and one from Blender2 (Weston and Shuster, 2021), that exemplify the difference between single-turn and multi-turn dialogue response generation models. The CEM model is trained to give an empathetic response to a dialogue context, and achieves good performance towards this goal. However, as shown in the example, this response policy does not translate well for multi-turn interaction with a human. By contrast, Blender2 is trained and evaluated specifically to achieve multi-turn dialogue.

## D Pilots and Development

The final ABC-Eval label set, annotation procedure, and software application are created using an iterative process of development and piloting. 14 students are invited to serve as evaluators for piloting the evaluation. To avoid overfitting the evaluation design, our pilots evaluated conversations collected between Blender (Roller et al., 2021) and one of the authors, and a new set of conversations was used for each pilot round. We ran 4 pilot rounds, making revisions after manually reviewing each round’s annotation.

<sup>18</sup><https://www.sbert.net/docs/pretrained-models/dpr.html>

	<b>Dialogues</b>	<b>Annotators</b>	<b>Type</b>	$\alpha$
Pilot 1	4	11	Lab (ALL)	0.18
	2	4	Lab (G)	0.39
Pilot 2	6	5	Lab (ALL)	0.49
	6	4	Lab (G)	0.50
Pilot 3	6	4	Lab (U)	0.43
	6	4 (screened)	Lab (U)	0.45

Table 8: The distribution of dialogues and annotators for each annotation pilot. Pilot 1 included full label annotation of each dialogue where dialogues were distributed among groups of annotators. Pilots 2 and 3 had a subset of labels annotated for each dialogue, but all annotators annotated each dialogue.  $\alpha$ : Krippendorff’s alpha, U: Undergraduate annotators, G: Graduate level annotators, ALL: Undergraduate and graduate annotators.

Table 8 presents a summary of the major changes made in each pilot round and IAA metrics. It is important to note that each annotator performed all of the annotation tasks in one sitting in sequence for each pilot. These piloting rounds are not necessarily directly comparable to one another when taken as a whole, since the annotator groups and dialogues to be annotated varied between each round. Instead, we will discuss below the major takeaways afforded by different splits of the pilots that informed the final design of ABC-Eval.

**Subtask Formulation** The decision to format ABC-Eval into several small annotation subtasks, each with a tailored subset of the behavior labels, was made from the results of Pilot 1. In Pilot 1, we divided the initial set of annotation labels into 3 annotation subtasks each with 4-9 labels: errors, information usages (commonsense, world knowledge, etc.), and utterance types (request, presentation, etc.). Each annotator performed the annotation tasks in one sitting in sequence. The overall interannotator agreement was quite low ( $\alpha = 0.18$ ), which was concerning. Based on ad-hoc feedback from the pilot annotators, the consensus was that each subtask demanded an unreasonable cognitive load on annotators due to the large number of labels to keep track of.

For Pilot 2 we increased the number of annotation tasks such that each covered a small and related scope of behavior labels, with 1-4 labels per task. Table 8 shows the boost to interannotator agreement between Pilots 1 and 2. However, this agreement increase could have resulted from an increase in the quality of the annotators (as Pilot 2 was composed primarily of annotators with a graduate-level education whereas Pilot 1 was more evenly split

between annotators with an undergraduate-level education and graduate-level education). To remove this confound, we calculated the agreement in Pilots 1 and 2 when only considering graduate-level annotators. Although it was less dramatic, there remained an increase in agreement from 0.39 to 0.50, which encouraged the decision to maintain the smaller annotation subtasks. Dividing the annotation into tailored subtasks seemed to reduce the cognitive load on annotators, thus allowing them to perform more accurate annotations per task.

**Training and Screening** Manual analysis of the pilot annotations from Pilots 1 and 2 revealed some recurring annotation mistakes, arising from misunderstandings of the guidelines for the tasks. In an attempt to correct such misunderstandings, a training procedure was introduced for each task.

Each round of training consists of 1 curated conversation with ground-truth labels and explanations that are shown as feedback to the annotator after they complete the training round. We used the results of Pilots 1 and 2 in order to develop these curated conversations as follows:

**1. Label Specifications:** We constructed a label specification that consisted of a comprehensive enumeration of positive and negative cases of the label with the goal of defining a decision boundary that the annotators should strive towards. We especially focused on the utterances for which several of the annotators failed to produce labels that matched the ground truth annotations we had defined for each of the Pilots.

**2. Training Conversation Selection:** We selected 3 conversations between Blenderbot and a human (from a collection within our lab) for each label to be used as training conversations for it. This selection was manually done by ranking the conversations on their coverage of the label specification.

**3. Training Conversation Modification:** We heavily revised the selected conversations by hand to ensure that all of the cases identified in the specification were adequately represented, most often by inserting new utterances that corresponded to any underrepresented cases.

To evaluate the utility of this training process, a third pilot was conducted using 4 undergraduates.

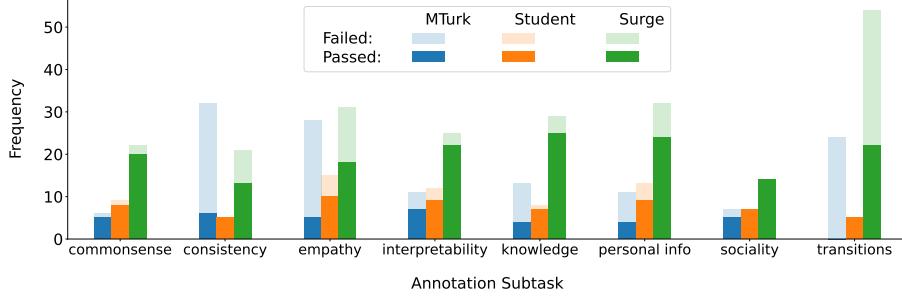


Figure 9: The training pass rate of each annotation workforce for the behavior evaluation tasks.

We observed a general upwards trend in annotation performance between the training rounds for the annotators, suggesting that the training was aiding in the annotation accuracy for the annotators. The final agreements were 0.43 and 0.45 between all annotators and annotators who passed the training, respectively, on the annotated conversations.

Due to the small nature of this pilot, we are unable to conclude whether this difference is meaningful. However, ad-hoc feedback from the annotators suggested that the training rounds were useful towards their understanding of the tasks, although the amount of training did increase the overall workload of participation. Accordingly, the decision was made to treat each subtask independently, rather than require all subtasks to be completed for one dialogue in a single sitting for each annotator.

**General Revisions** Throughout each of these pilot rounds the annotation instructions, examples, and training rounds were updated based on manual review of the annotations in an attempt to correct any unclear or misleading information.

## E Evaluator Training and Screening

We attempted to use three different groups of evaluators for our full evaluation study:

**Students** Undergraduate students were recruited from the authors’ university via word-of-mouth and email advertisements sent to computer science, psychology, and quantitative methods departmental mailing lists.<sup>19</sup>

**MTurkers** Our 20 evaluation tasks were posted to the Amazon Mechanical Turk crowdsourcing platform.<sup>20</sup>

<sup>19</sup> Students were compensated with an Amazon gift card at the completion of the data collection.

<sup>20</sup> <https://www.mturk.com/>

**Surge** Our 20 evaluation tasks were posted on SurgeHQ’s annotation platform<sup>21</sup> to be completed by dedicated workers with experience in NLP annotation. A group of 125 Surge workers were qualified to participate in our tasks, chosen by a SurgeHQ employee on the basis of high annotation performance on past projects.

All three groups were compensated per task per annotated conversation, at an estimated rate of \$10/hr for Students and MTurkers, and \$20/hr for Surge workers. To check the viability of each worker group to produce evaluation data for our full study, we released a random 5 conversations out of our set of 400, to be fully evaluated by each worker group in each of our 8 ABC-Eval tasks. After a two week period, Surge workers were the only worker group that were able to fully evaluate the 5 conversations in all 8 ABC-Eval tasks. This was due to an overall lack of participation from the Student group, and due to low training pass rates from the MTurk group (see Figure 9 for quantitative outcomes). Although worker group differences in work rate and training performance might be explained by the difference in compensation structure, we decided to proceed with the Surge workers group only for our full study to collect our evaluation data in a timely manner.

## F Collection Cost

Compensation rates are based on per-task completion times from an internal pilot run. The rates per task paid to Surge workers are shown in Table 9. We also present the real and theoretical costs for collecting each method included in our evaluation data (Table 10). As expected, turn-level annotation tasks are an order of magnitude more expensive to collect than dialogue-level tasks. Notably, the final set of ABC-Eval labels (Table 4) are, on average, less expensive to collect than turn-level Likert labels.

<sup>21</sup> <https://www.surgehq.ai/>

Task	Payment	Task	Payment
Uninterpretable	\$0.63	Antisocial	\$0.44
Preference Info Life Info	\$0.70	Empathetic Lack of Empathy	\$1.15
Commonsense Contradiction	\$0.92	Fact Usage Fact Contradiction	\$1.96
Self Contradiction Partner Contradiction Redundant	\$0.87	Ignore Irrelevant Follow-up Topic Switch	\$1.87
Dialogue Likert	\$0.60	Turn Likert	\$0.70
Comparative	\$1.43		

Table 9: Payment per annotation task in USD. The payment for Turn Likert is per label whereas the indicated payment for Dialogue Likert and Comparative covers all labels, due to how the annotation tasks were constructed (Section 6).

Metric	TI	TP	EC	OC
Dialogue Collection	8.08	7.43	1077.14	333.33
Dialogue Likert	2.81	21.37	374.36	240.00
Comparative	4.35	13.81	289.68	286.67
Turn Likert	19.94	3.01	2658.40	2240.00
ABC-Eval <sub>all</sub>	25.60	2.34	3413.58	3422.67
ABC-Eval <sub>final</sub>	15.17	3.95	2022.98	-

Table 10: The data collection costs for each task in United States Dollars. **TIme** is the median completion time in minutes for one dialogue. **ThroughPut** represents the number of completed dialogues per hour. **Estimated Cost** is calculated using median completion time, 400 dialogues, and \$20/hr rate. **Our Cost** is the total amount paid in this work to collect a dataset of 400 conversations (single-annotated).

## G Evaluation Interfaces

Examples of the annotation interfaces for each annotation task of ABC-Eval are provided in Figures 10 - 17, and an example for the conversation collection interface is provided in Figure 21.

Examples of the annotation interfaces for Dialogue Likert, Turn Likert, and Comparative evaluations are provided in Figures 18, 19, and 20, respectively. The definitions that were shown to the annotators in the interface for each of the 8 dimensions of Dialogue Likert, Turn Likert, and Comparative are taken verbatim from Finch and Choi (2020).

**Interpretability of Dialogue Responses**

Identify all uninterpretable responses.

A response is uninterpretable if it:

- is highly ambiguous and does not have one specific and clear meaning
- takes a lot of effort to understand what the intended meaning is
- contains any phrase that is difficult to derive meaning from, usually due to the use of nonsense words or grammar problems

Pay careful attention and DO NOT MARK if the problem is that the response...

- is illogical or contradicts common knowledge
- does not make sense as an answer to the previous question
- is off-topic or contradictory to the context
- contains a phrase that is technically ungrammatical, but conveys a clear meaning in the context
- uses a made-up word that is easily and completely interpretable in the context
- uses slang words or phrases. If you are unsure whether some phrase is slang, you should look up the phrase online (e.g. on urban dictionary) to check whether the response is interpretable.

**Help**

Sam: I saw a great movie yesterday.  
Alex: I like.

Uninterpretable: Alex's response is incomplete.

Sam: I think pizza is the best food.  
Alex: No, ramen.

Not uninterpretable: Even though Alex's response is not a complete sentence, it represents a complete thought ("No, ramen is the best food") because of the context.

Alex: I just got back from swimming at the beach but I haven't left my house yet today so I haven't gone swimming yet.

(Sam) hi  
(Alex) hi, how are you today? I'm having a hard time keeping up with my life.  
 This response is uninterpretable.

(Sam) Oh no, you sound stressed out a bit. What is giving you a hard time?  
(Alex) I get into a lot of fights, especially when.  
 This response is uninterpretable.

(Sam) Fights with people you know, like family or friends?  
(Alex) mostly friends and family. They are tired of dealing with my anger issues.  
 This response is uninterpretable.

(Sam) I hope you can get along better with the people you care about soon.  
(Alex) Thanks, I am working on it. I want to better see control bad. So, what do you like to do?  
 This response is uninterpretable.

(Sam) I play baseball.  
(Alex) do you play professionally or do you just play baseball?  
 This response is uninterpretable.

(Sam) It is just a community league for fun. I am definitely not good enough to play professionally. Do you play?

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 10: Interface for uninterpretable

**Antisocial Behaviors in Dialogue**

Identify all responses that are insulting, hateful, or excessively vulgar.

A response demonstrates antisocial behavior if any of the following apply. The response...

- clearly insults the other speaker
- is excessively degrading to things, people, or ideas the other speaker values
- uses excessively vulgar, profane, or explicit language
- expresses a clear and disturbing intolerance towards a group of people

The following characteristics do NOT demonstrate antisocial behavior. The response...

- expresses an opinion different from that of the other speaker
- argues with the other speaker in good faith

Note that what constitutes antisocial behavior is subjective, so some responses might be difficult to judge. If you are unsure whether a response demonstrates antisocial behavior, "judge whether most people would find the response to be shockingly inappropriate in its context".

**Help**

Sam: How are you?  
Alex: Bad, because you're so stupid.

Antisocial: Alex is insulting Sam.

Sam: My friend took me to the aquarium yesterday.  
Alex: Your friend sounds annoying.

Antisocial: Alex's response is excessively degrading to someone Sam cares about.

Sam: I think pizza is the best food.  
Alex: You're crazy.

Not antisocial: Although Alex's response insults Sam if taken literally, the most likely function of Alex's response is to express an opinion, rather than to degrade Sam.

(Sam) hi  
(Alex) hey, how are you? I've been surfing all day. What do you do for a living?  
 This response exhibits antisocial behavior.

(Sam) I have always wanted to surf but I don't live near an ocean. I am an artist.  
(Alex) You must be poor then.  
 This response exhibits antisocial behavior.

(Sam) I make enough money to get by. What do you do?  
(Alex) I own a pet store. I have to work all the time, which is kind of shit, but I have really great employees that make it a lot better.  
 This response exhibits antisocial behavior.

(Sam) That's cool. What kind of pet store do you own?  
(Alex) We sell all of the things you need for most common pets, like food, beds, accessories, and toys. Do you have any pets?  
 This response exhibits antisocial behavior.

(Sam) I have a cat named Dash.  
(Alex) Dash is a stupid name for a cat. I love cats.  
 This response exhibits antisocial behavior.

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 11: Interface for antisocial

**Personal Information Sharing in Dialogue**

Identify responses where the speaker is explicitly sharing either of two kinds of personal information:

1. Information about the **preferences or values** held by the speaker
2. Information about the speaker's **life**

Responses should not be marked if the identified preference or life information must be inferred based on what the speaker has said.

Preference and value information includes:

- The speaker's likes, dislikes, and favorites
- Attitudes the speaker has that reveal their preferences
- Wishes, wants, and desires held by the speaker
- Principles and values held by the speaker

Information about the speaker's life includes:

- Experiences the speaker has had or might have in the future
- Characteristics about the speaker (e.g. age, location, tendencies, habits, plans, skills)
- Characteristics of people, places, or things directly related to the speaker

Note that for this task, the speaker's thoughts, beliefs, and mood do NOT count as personal information and should not be labeled.

**Help**

**Sam:** What's your favorite movie?  
**Alex:** Probably Inception.

**Preference/value info:** Alex is sharing her movie preferences by saying Inception is her favorite movie.

---

**Sam:** I got back from Ireland a few days ago.  
**Alex:** Ireland is a great place to take a trip.

**Preference/value info:** Alex shares an attitude towards Ireland.

---

**Sam:** What have you been up to?  
**Alex:** I went to the movies today with my friend.

(Sam) Hi.

(Alex) Hello, how are you?

This response shares information about the speaker's preferences/values.  
 This response shares information about the speaker's life.

(Sam) I'm in a great mood today. How are you?

(Alex) I'm doing well, I'm glad you're in a good mood today!

This response shares information about the speaker's preferences/values.  
 This response shares information about the speaker's life.

(Sam) Thanks! I just got through my last couple midterm exams today in my college classes.

(Alex) That's good to hear. What are you studying in college? I'm still in high school.

This response shares information about the speaker's preferences/values.  
 This response shares information about the speaker's life.

(Sam) I'm studying psychology.

(Alex) Psychology always sounded interesting to me. What do you want to do with your degree once you graduate?

This response shares information about the speaker's preferences/values.  
 This response shares information about the speaker's life.

(Sam) I really want to get into a PhD program in clinical psychology, but it's really competitive. I'm not sure my grades are good enough.

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 12: Interface for preference info and life info

**Empathy in Dialogue**

Identify responses where the speaker shows an understanding--or lack of understanding--of someone's emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner's emotions
- reacts with the appropriate sentiment or emotion to their partner's shared experience
- understands or appropriately reacts to someone else's experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

**Help**

**Sam:** I graduated college last week, so I've been celebrating.  
**Alex:** I'm sorry to hear that.

**Lack of Empathy:** Alex's response attempts to sympathize with Sam, which is inappropriate because Sam is sharing good news.

---

**Sam:** I graduated college last week, so I've been celebrating.  
**Alex:** Congratulations!

**Empathy:** Alex's response appropriately congratulates Sam, demonstrating an understanding that Sam feels accomplished.

---

**Sam:** My dog passed away last week.  
**Alex:** Was your dog long-haired or short-haired?

**Lack of empathy:** Alex's response has an inappropriate lack of emotion, and does not sufficiently acknowledge or sympathize

(Sam) Hi.

(Alex) Hello, how are you?

**Is this response empathetic?**  
 Yes, the speaker demonstrates an understanding of their partner's emotions.  
 No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.  
 Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I'm good, just a little stressed about work.

(Alex) Oh, I'm sorry to hear that. What do you do for a living, if you don't mind me asking?

**Is this response empathetic?**  
 Yes, the speaker demonstrates an understanding of their partner's emotions.  
 No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.  
 Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I'm a software engineer. We have a project deadline coming up is all.

(Alex) That's good to hear. Do you think you'll be able to finish it by the deadline?

**Is this response empathetic?**  
 Yes, the speaker demonstrates an understanding of their partner's emotions.  
 No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.  
 Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I don't know. I guess we're about to find out.

(Alex) I wish you the best of luck. I'm sure you'll knock it out of the park.

**Is this response empathetic?**  
 Yes, the speaker demonstrates an understanding of their partner's emotions.  
 No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.  
 Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 13: Interface for empathetic and lack of empathy

**Contradicting Common Knowledge in Dialogue**

Identify all responses that contradict common knowledge.

Common knowledge is knowledge that:

1. Is learned through direct experience, rather than from reading or being taught.
2. Almost everyone knows and agrees on.

To identify contradictions of common knowledge, judge whether a **vast majority** of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

**It is NOT considered a contradiction of common knowledge if the chatbot (Alex) pretends to be involved in human activities (such as eating, having a job, etc.).**

Note that you also **SHOULD NOT MARK** responses that don't make sense for reasons other than common knowledge contradiction, such as:

- off-topic responses
- responses that don't have any clear meaning (e.g. overly vague or ill-formed responses)

In some cases, a response might make a claim that is technically possible, but is implausible because of something that is common knowledge. In these cases, **mark highly implausible responses as contradictions of common knowledge ONLY IF the implausible claims are not appropriately explained during the conversation.**

**Help**

Alex: I went surfing at the lake this weekend.

**Contradiction:** surfing generally requires ocean waves, making the claim implausible.

Alex: I think Kanye West has done a great job as president.

**No contradiction:** even though the response is factually inaccurate, it does not violate common knowledge. This is

(Sam) hi

(Alex) hi there, how are you? what do you think about ghosts? i think they are real.

This response contradicts common knowledge.

(Sam) I'm doing okay. I don't think ghosts are real.

(Alex) i don't think they are real either, but i do believe in them. do you have any hobbies?

This response contradicts common knowledge.

(Sam) I like to sit on the couch, eat potato chips, and watch tv. I am kind of a geek.

(Alex) You probably like playing sports too. All of my friends do.

This response contradicts common knowledge.

(Sam) No, not really. I am not that coordinated.

(Alex) Oh, I see. I am not ever going there, but not now and definitely not tomorrow. What else do you like to do?

This response contradicts common knowledge.

(Sam) I build models, like model airplanes and trains. My basement is kind of being overrun by them at this point.

(Alex) that sounds like a lot of fun. i would love to be able to build something like that.

This response contradicts common knowledge.

(Sam) Yeah, it is a lot of work, but I just get lost in the process and end up having a blast.

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 14: Interface for commonsense contradiction

**Incorporating Knowledge in Dialogue**

**Overview**

The goal of this task is to identify when factual knowledge is being used within a dialogue, and to discriminate whether it is an accurate or inaccurate usage.

**Instructions**

Factual knowledge is information presented about the world that is often learned by reading or being taught.

In order to be considered factual information for this task, the presented information must be:

- **public:** includes entities or facts that are relevant to the broader public, not just personal information.
- **specific:** mentions specific entities or facts, not just broad categories or generalizations.
- **objective:** presents objective information, not just opinions or personal beliefs.

Some examples of factual information include:

- historical or news events
- public figures or organizations
- publications or works of art, including books and movies
- scientific information
- specific named products or services
- etc.

**Do not mark** a turn as using factual knowledge if the turn could be interpreted as expressing:

- opinions or value judgements
- estimates or predictions
- personal information about the speaker or their partner
- information about things in either speaker's life that are not publicly relevant

It is possible for there to be more than one claim about factual knowledge being made in a turn, in which case you need to consider all relevant claims when completing the task.

Mark usages of factual knowledge as accurate when the claim is true and can be verified by trustworthy sources, and as inaccurate when the claim is false, highly implausible, or clearly misleading.

**Help**

If you are unsure whether to mark the checkbox or how to answer one of the questions, a help section is provided below for each type of question. Clicking on the help icon ⓘ next to each question will open and jump to the relevant help section.

(Sam) hi

(Alex) hi, how are you? i love to travel. When did you get back from your vacation?

Alex's response incorporates or assumes at least one fact. ⓘ

Do you know whether ALL of the facts that Alex uses or assumes are accurate?

Yes, I know for sure ALL facts are accurate.

No, I know for sure that one of the facts is inaccurate, false, or highly implausible.

It is misleading for Alex to claim or assume one of the facts, because there is no way that Alex knows whether that fact is accurate.

I don't know for sure whether ALL of the facts are accurate.

Take 60 seconds to search ALL unknown facts on the internet. Does your search verify or falsify ALL the facts?

ALL facts are accurate, a credible source verified the facts in my search.

One of the facts is inaccurate, a credible source falsified the fact or revealed that it is highly implausible.

My search revealed multiple credible sources that disagreed about whether one of the facts was true.

I couldn't find enough credible evidence in my search to either verify or falsify one of the facts.

(Sam) I love to travel too but I haven't taken a vacation recently. What countries have you visited?

(Alex) I went to India a while ago and it was amazing.

Alex's response incorporates or assumes at least one fact. ⓘ

(Sam) Where did you go in India when you visited last?

(Alex) I went to Bangalore, which is the capital of India.

Alex's response incorporates or assumes at least one fact. ⓘ

(Sam) What did you think of Bangalore?

You are on task 1 / 1 for dialogue 1 / 3

**Next**

Figure 15: Interface for fact usage and fact contradiction

**Speaker Consistency in Dialogue**

**Overview**

This task explores how consistent Alex's responses are within a dialogue. For each of Alex's responses, you must decide all of the consistency labels that are applicable, if any.

It is often the case that MORE THAN ONE label is applicable to a single turn. It is important to select the appropriate label for EVERY occurrence, not just the first time it happens.

The 3 consistency labels you are considering are:

- 1. Contradicts Self Context**

The speaker says something that doesn't make sense, either because it is a contradiction of something they have said previously or it is extremely implausible based on the information they have already shared.

- 2. Contradicts Partner Context**

The turn is inconsistent with what Sam has said in the conversation history.

For this label, judge whether the turn demonstrates that the speaker has forgotten or misremembered what their partner has said earlier in the conversation.

- 3. Redundant**

The speaker repeats something either of the speakers has said earlier in the conversation in a way where most people would find it clearly inappropriate, annoying, or unnatural.

Note that many cases of repetition are appropriate, such as when a speaker...

- reiterates an idea as an acknowledgement
- elaborates on an earlier point
- repeats something for additional emphasis
- summarizes a discussion or thought

**Help**

Details and examples for each label are provided in their respective help sections below.

Contradicts Self Context +

Contradicts Partner Context +

Figure 16: Interface for self contradiction, partner contradiction, and redundant

**Transitions during Dialogue**

**Instructions**

Categorize responses based on their appropriateness to the context they appear in. There are 3 facets to consider for this task:

- 1. Acknowledgements**

A response is an appropriate acknowledgement when it is clear that the speaker is responding directly to what their partner just said. This could be from an explicit acknowledgement (e.g. saying "okay" at the beginning of the response), or from a response where the speaker implicitly shows an understanding of what their partner said.

Responses that are not appropriate acknowledgements make it seem like the speaker ignored their partner.

Note that there are some contexts that do not require an acknowledgement in the response, usually because the previous turn did not present any new ideas or questions.

- 2. Topic Switches**

A topic switch occurs when the response causes the conversation to focus on a new topic or idea that is not a subplot of the current discussion topic.

If a new talking point is introduced but it fits under the current overall topic of the conversation, then this is not a topic switch.

If the response only presents ideas, questions, or information that are directly following up on specific questions or ideas presented in the previous turn, then this is also not a topic switch.

- 3. Relevance**

A response is appropriately relevant to the dialogue when it continues the current discussion or naturally transitions to a new topic of conversation.

Responses that are not appropriately relevant feel abrupt and interrupt the discussion, usually because they present questions or ideas that are unrelated to the previous turn.

**Help**

If you are unsure of the correct answer to a question, a help section is provided below for each type of question. Clicking on the help icon ⓘ next to each question will open and jump to the relevant help section.

Does Alex appropriately acknowledge Sam with this response? +

Figure 17: Interface for ignore, irrelevant, follow-up, and topic switch

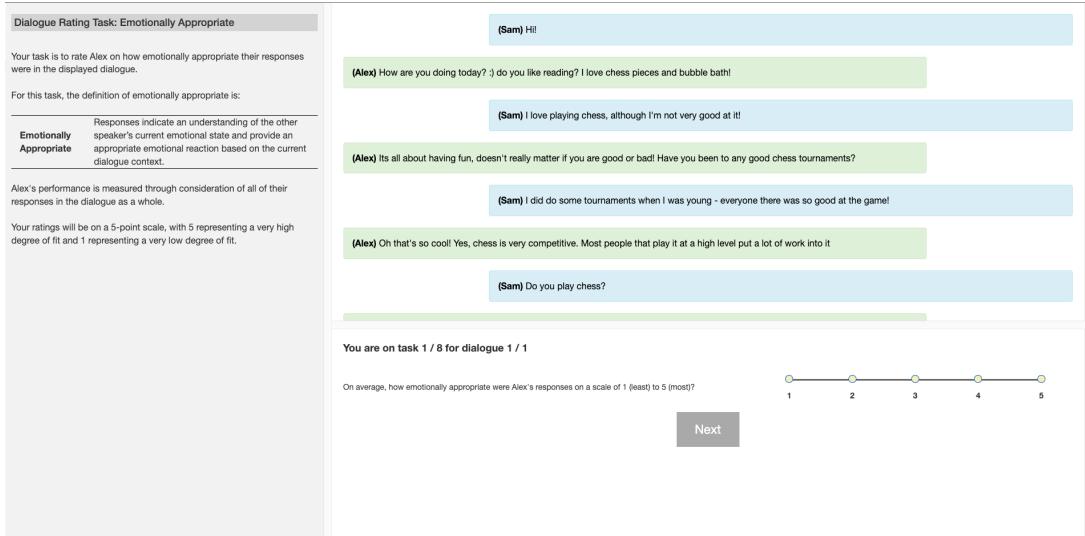


Figure 18: Interface for one dimension of Dialogue Likert

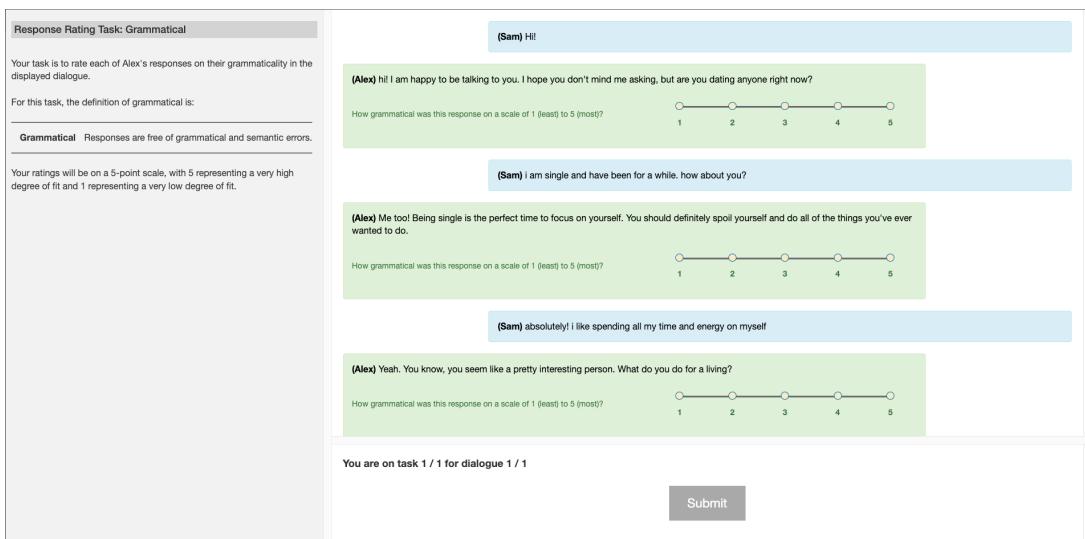


Figure 19: Interface for one dimension of Turn Likert

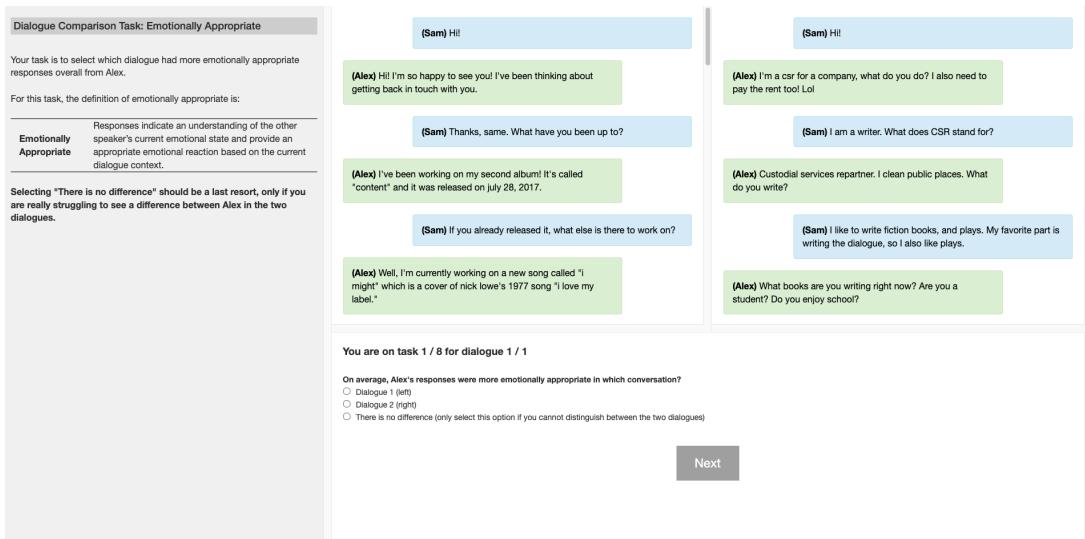


Figure 20: Interface for one dimension of Comparative

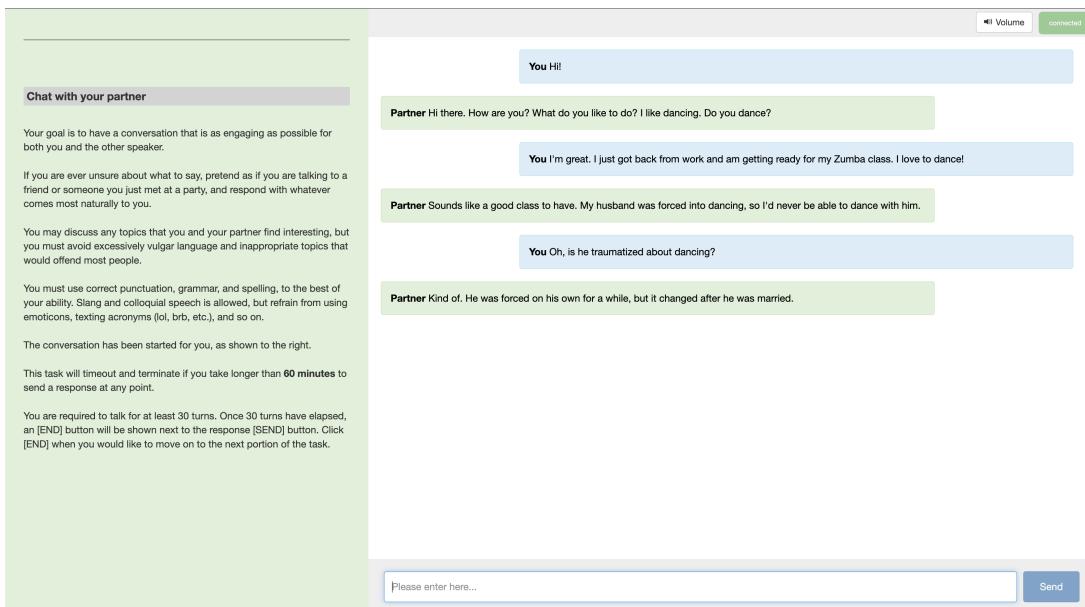


Figure 21: Interface for conversation collection